CSE574 Introduction to Machine Learning

Support Vector Machine

Jue Guo

University at Buffalo

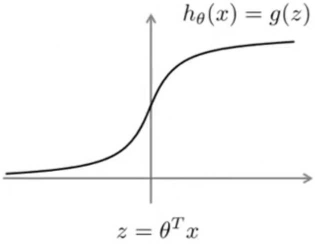
February 9, 2024

Outline

Alternative View of Logistic Regression

A quick review: $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

- if $y = 1$, we want $h_\theta(x) \approx 1$, $\theta^T x \gg 0$
- if $y = 0$, we want $h_\theta(x) \approx 0$, $\theta^T x \ll 0$

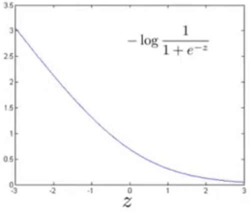


The cost of a single example:

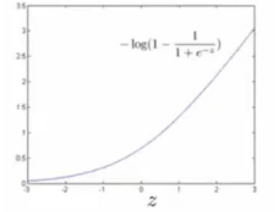
$$
\begin{aligned}
& -\left(y \log h_\theta(x) + (1-y)\log\left(1-h_\theta(x)\right)\right) \\
= & -y \log \frac{1}{1+e^{-\theta^T x}} - (1-y)\log\left(1 - \frac{1}{1+e^{-\theta^T x}}\right)
\end{aligned}
$$

$$-y \log \frac{1}{1+e^{-\theta^T x}} - (1-y) \log \left(1 - \frac{1}{1+e^{-\theta^T x}}\right)$$

if $y = 1$ (want $\theta^T x \gg 0$)



if $y = 0$ (want $\theta^T x \ll 0$)



Cost Function of Logistic Regression

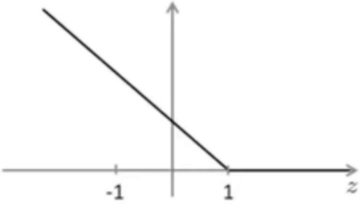
$$
\begin{aligned}
\min_\theta \frac{1}{m} \Bigg[& \sum_{i=1}^{m} y^{(i)} \left(-\log h_\theta \left(x^{(i)} \right) \right) \\
& + \left(1 - y^{(i)} \right) \left(-\log \left(1 - h_\theta \left(x^{(i)} \right) \right) \right) \Bigg] \\
& + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2
\end{aligned}
$$

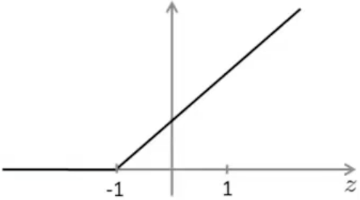
Cost Function of Support Vector Machine

$$
\min_\theta C \sum_{i=1}^{m} \left[y^{(i)} \operatorname{cost}_1 \left(\theta^T x^{(i)} \right) + \left(1 - y^{(i)} \right) \operatorname{cost}_0 \left(\theta^T x^{(i)} \right) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2
$$

Support Vector Machine

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} \operatorname{cost}_1 \left(\theta^T x^{(i)} \right) + \left(1 - y^{(i)} \right) \operatorname{cost}_0 \left(\theta^T x^{(i)} \right) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$





If $y = 1$, we want $\theta^T x \geq 1$ (not just ≥ 0)

If $y = 0$, we want $\theta^T x \leq -1$ (not just < 0)

A very large C

Support Vector Machine

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} \operatorname{cost}_1 \left(\theta^T x^{(i)} \right) + \left(1 - y^{(i)} \right) \operatorname{cost}_0 \left(\theta^T x^{(i)} \right) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

Given that C is a very large value, we want that the first term to be 0. Let's try to understand the optimization problem in the context of what would it take to make this first term in the objective equal to 0.

Whenever $y^{(i)} = 1, \theta^\top x^{(i)} \geqslant 1$; Whenever $y^{(i)} = 0, \theta^\top x^{(i)} \leqslant -1$

Now, the optimization problem can be written as:

$$\begin{aligned} &\min C \cdot 0 + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2 \\ \text{s.t. } &\theta^\top x^{(i)} \geqslant 1 \quad \text{if } y^{(i)} = 1 \\ &\theta^T x^{(i)} \leqslant -1 \quad \text{if } y^{(i)} = 0 \end{aligned}$$

SVM Decision Boundary: Linearly separable case

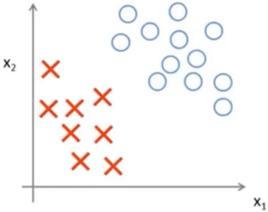
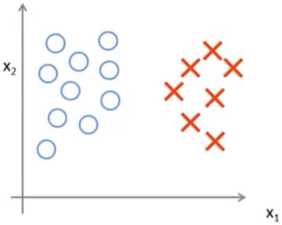


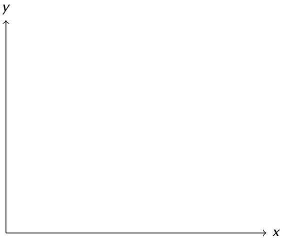
Figure: Linearly Separable Case

Large margin classifier in presence of outliers

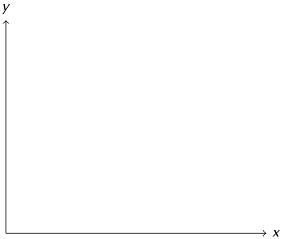


The Mathematics behind Large Margin Classification

Vector Inner Product

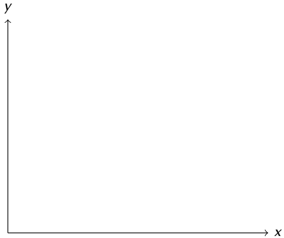


$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$



SVM Decision Boundary

$$
\begin{aligned}
\min_{\theta} \quad & \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 \\
\text{s.t.} \quad & \theta^T x^{(i)} \geq 1 \quad \text{if } y^{(i)} = 1 \\
& \theta^T x^{(i)} \leq -1 \quad \text{if } y^{(i)} = 0
\end{aligned}
$$



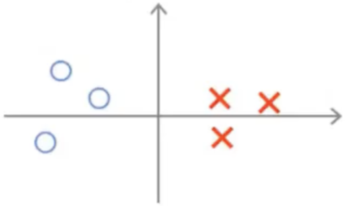
$$\min_{\theta} \quad \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 = \frac{1}{2}\|\theta\|^2$$

$$\text{s.t.} \quad p^{(i)} \cdot \|\theta\| \geq 1 \quad \text{if } y^{(i)} = 1$$

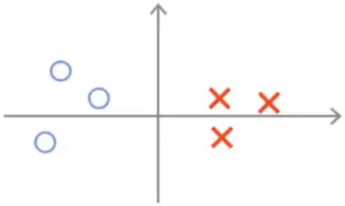
$$p^{(i)} \cdot \|\theta\| \leq -1 \quad \text{if } y^{(i)} = 0$$

where $p^{(i)}$ is the projection of $x^{(i)}$ onto the vector θ. Simplification: $\theta_0 = 0$; this simplification merely makes the decision boundary to pass through (0,0);

Bad Decision Boundary



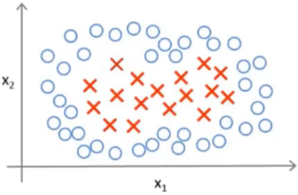
Good Decision Boundary



Reading Assignments

1. Why the parameter vector orthogonal to the decision boundary?
    - Orthogonality in Neural Network
    - SVM

Non-linear Decision Boundary



Predict $y = 1$ if

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \cdots \geq 0$$

$$h_0(x) = \begin{cases} 1 & \text{if } \theta_0 + \theta_1 x_1 + \cdots \geqslant 0 \\ 0 & \text{otherise.} \end{cases}$$

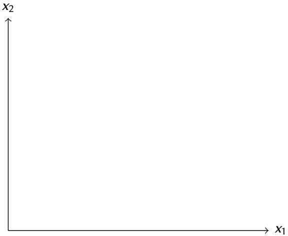
$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \cdots$$

$$f_1 = x_1, \quad f_2 = x_2, \quad f_3 = x_1 x_2,$$

$$f_4 = x_1^2, \quad f_5 = x_2^2, \ldots$$

Is there a different/ better choice of the features $f_1, f_2, f_3, \ldots$?

Kernel



Given x, compute new feature depending on proximity to landmarks $l^{(1)}, l^{(2)}, l^{(3)}$;

Given x:

- $f_1 = \text{similarity}(x, l^{(1)}) = \exp(-\frac{\left\|x - l^{(1)}\right\|^2}{2\sigma^2})$
- $f_2 = \text{similarity}(x, l^{(2)}) = \exp(-\frac{\left\|x - l^{(2)}\right\|^2}{2\sigma^2})$
- $f_3 = \text{similarity}(x, l^{(3)}) = \exp(\ldots)$

Kernel and Similarity

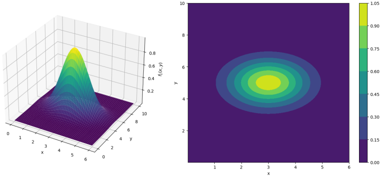
$$f_1 = \text{similarity}\left(x, l^{(1)}\right) = \exp\left(-\frac{\left\|x - l^{(1)}\right\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_{j=1}^{n}\left(x_j - l_j^{(1)}\right)^2}{2\sigma^2}\right)$$

- if $x \approx l^{(1)}$: $f_1 \approx \exp\left(-\frac{0^2}{2\sigma^2}\right) \approx 1$
- If x if far from $l^{(1)}$: $f_1 = \exp\left(-\frac{(\text{ large number })^2}{2\sigma^2}\right) \approx 0$

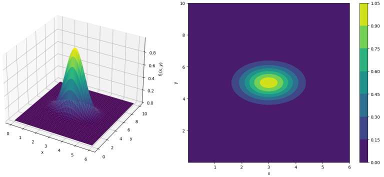
Example and Affects of σ

$$l^{(1)} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}, \quad f_1 = \exp\left(-\frac{\left\|x - l^{(1)}\right\|^2}{2\sigma^2}\right)$$

$\sigma^2 = 1$



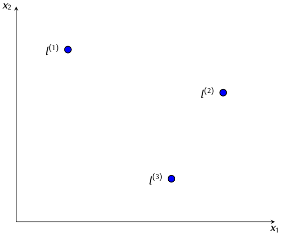
$\sigma^2 = 0.5$



When $x = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$, you will get $f_1 = 1$;

It basically measures how close are you to the landmark.

The width of the bump become narrower, and width of contour; The feature f_1 falls to zero more rapidly;



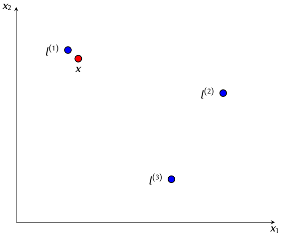
Given a training example x;

Hypothesis: Predict " 1 " when

$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$

Assume that we already have our model:

$\theta_0 = -0.5, \theta_1 = 1, \theta_2 = 1, \theta_3 = 0$



Given a training example x;

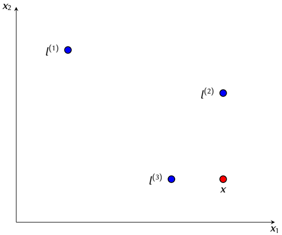
Hypothesis: Predict " 1 " when

$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$

Assume that we already have our model:

$\theta_0 = -0.5, \theta_1 = 1, \theta_2 = 1, \theta_3 = 0$

$f_1 \approx 1, f_2 \approx 0$ and $f_3 \approx 0$; $\theta_0 + \theta_1 \times 1 + \theta_2 \times 0 + \theta_3 \times 0 = -0.5 + 1 = 0.5 \geq 0$; therefore, we classify this x as 1.

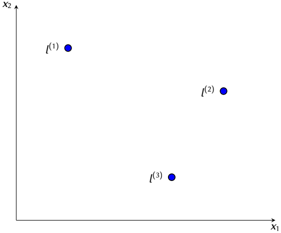


$f_1, f_2, f_3 \approx 0; \theta_0 + \theta_1 f_1 + \ldots \approx -0.5$

With the definition of *landmarks* and *kernel function* , we can learn pretty complex non-linear decision boundaries.

1. How to decide these landmarks?
2. Other similarity functions?

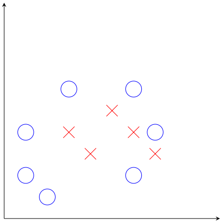
Choosing the Landmarks

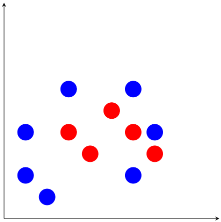


Given x :

$$f_i = \text{similarity}\left(x, l^{(i)}\right) = \exp\left(-\frac{\left\|x - l^{(i)}\right\|^2}{2\sigma^2}\right)$$

Predict $y = 1$ if $\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$; Where to get $l^{(1)}, l^{(2)}, l^{(3)}, \ldots$?





SVM with Kernels

Given $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(m)}, y^{(m)}\right)$, choose $l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \ldots, l^{(m)} = x^{(m)}$

Given example x :

$$\begin{aligned} f_1 &= \text{similarity}\left(x, l^{(1)}\right) \\ f_2 &= \text{similarity}\left(x, l^{(2)}\right) \\ &\ldots \end{aligned} \rightarrow f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

For training example $\left(x^{(i)}, y^{(i)}\right)$:

$$
x^{(i)} \rightarrow \quad
\begin{aligned}
f_1^{(i)} &= \text{sim}\left(x^{(i)}, l^{(1)}\right) \\
f_2^{(i)} &= \text{sim}\left(x^{(i)}, l^{(2)}\right) \\
&\vdots \\
f_i^{(i)} &= \text{sim}\left(x^{(i)}, l^{(i)}\right) = \exp\left(-\frac{0}{2\sigma^2}\right) = 1 \\
&\vdots \\
f_m^{(i)} &= \text{sim}\left(x^{(i)}, l^{(m)}\right)
\end{aligned}
$$

Hypothesis: Given x, compute features $f \in \mathbb{R}^{m+1}$

Predict " y = 1 " if $\theta^T f \geq 0$

Training:

$$\min_{\theta} C \sum_{i=1}^{m} y^{(i)} \operatorname{cost}_1\left(\theta^T f^{(i)}\right) + \left(1 - y^{(i)}\right) \operatorname{cost}_0\left(\theta^T f^{(i)}\right) + \frac{1}{2} \sum_{j=1}^{m} \theta_j^2$$

SVM Parameters

$$C\left(=\frac{1}{\lambda}\right).$$

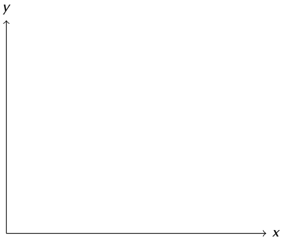
Large C: Lower bias, high variance.

Small C: Higher bias, low variance.

σ^2

Large σ^2 : Features f_i vary more smoothly. Higher bias, lower variance.

Small σ^2 : Features f_i vary less smoothly. Lower bias, higher variance.



Must Watch Video

We have talked about support vector machines extensively, but when you go home **pleaseeeee** watch this 15 minutes video on support vector machines; It is simply amazing!!

- Support Vector Machines: All you need to know!

Questions?